ANANSI

Premises

General Scope

This project brief relates to the CLARIAH Technical Plan for Work Package 2. The principle goal for CLARIAH WP2 is the construction of the overarching components of the individual CLARIAH research infrastructures - which are the subject of WP3: Linguistics, WP4: Social & Economical History & WP5: Media Studies.

Project Scope

In the project a number of tasks defined for CLARIAH WP2 in the Technical Plan are combined and executed as a single project with the name ANANSI. The relevant tasks are:

- 11.100 - RDF Infrastructure
- 11.200 - Maintenance Tools
- 11.300 - Data Conversion Tools
- 41.100 - Data Access Environment
- 42.100 - Software plugins

Proposition

Problem

The humanities developed many authoritative data and research infrastructures over the last decade. These infrastructures often cover single scholarly fields but contain data valuable for other fields and/or research methods. Since the data models are frequently customised the interoperability between these infrastructures is limited. This problem is expanded by differing technical standards, access limitations because of commercial and intellectual property rights etc. All of these problems make it inefficient or even infeasible to combine multiple data sources in order to identify overlap or to expand the use of the research data beyond the original scope.

Goal

ANANSI will first of all form the data hub between the three primary CLARIAH domains (Linguistics, Social & Economical History and Media Studies). This hub will facilitate the connectivity of structured (meta-)data served by the research infrastructures constructed in respectively WP3, 4 and 5. Secondly, ANANSI will integrate the data with large scale existing data infrastructures outside CLARIAH - both within and outside the humanities.

Solution

ANANSI will provide a Linked Open Data (LOD) infrastructure that links to the original research data, which remains within the field for which it was created and modelled. Data entities served through ANANSI will be as “indisputable” as possible. The core entities answer the *who, what* and *where* questions: [Persons], [Concepts], [Locations]. At an earlier stage the *when* question [Events] has been discussed but this has currently been placed on hold.

ANANSI harvests a limited subset of attributes from the data providing infrastructures to populate the core entities. The primary function of these attributes is to uniquely identify the core entity - e.g. a [Person] entity may have a [names], [birth date], [birth location], [death date], [death location] and [gender] property. All data entities present in ANANSI are represented in commonly used and well documented standards.

ANANSI will provide both a machine readable web-API and a human readable front-end environment to access, query, manage and convert data. The machine readable web-API will be demonstrated by creating connections to ANANSI from a number of software systems frequently used for (digital) humanities research.

ANANSI must be fully open access - the limited subset of attributes of all data entities will be provided to all visitors and requests. The original data behind the connecting links may be subject to access limitations.

Maintenance of the data in ANANSI is the primary responsibility of the providers. Data harvested will be provided as-is with a provenance track to the original source. For deduplication purposes automated tools must be implemented. The providers are responsible for providing the attributes in a format/vocabulary that allows semantic integration. There is no organised human board of editors planned.

Target

Audience

The general audience of CLARIAH are humanities researchers who are served through the work packages 3-Linguistics, 4-Social & Economic History and 5-Media Studies.

The audience of ANANSI is defined by work package 2. Work package 2 considers its primary audience as software engineers and computer scientists working within the fields of language processing and digital humanities. These engineers are mostly employed by universities, libraries, archives and research facilities who engage directly with the humanities faculties and cultural heritage institutions.

The group of humanities scholars with (advanced) computational skills is rapidly growing. This is clearly visible in the rising attendee levels at the ADHO DH20xx conferences, stimulated by EU projects like Dixit and the DH Benelux conferences. Collectively known as digital humanities scholars, this group forms work package 2’s secondary audience.

Scope

ANANSI aims to provide data relevant for the Humanities, covering the Netherlands, its inhabitants and its language in its widest form. Primary focus is on the modern state of the Netherlands. Secondary focus is on the borders of the modern Dutch language, which expands the scope to Flanders, Belgium. Third focus is on the historical region known as the Low Countries which widens the scope to the whole of (French and German speaking) Belgium, Luxemburg, parts of northern France and western Germany. Fourth focus lies with the historical Dutch colonies, settlements and trading posts all over the world including the three other constituent countries within the current Kingdom of the Netherlands - Aruba, Curaçao and St. Martin - and the overseas regions of Bonaire, St. Eustatius and Saba.

Metrics

Criteria

ANANSI will be considered successful when the following criteria are met:

1. ANANSI is connected to a WP3 LOD-enabled data repository and serves a (selected) representation of its data;
2. ANANSI is connected to a WP4 LOD-enabled data repository and serves a (selected) representation of its data;
3. ANANSI is connected to a WP5 LOD-enabled data repository and serves a (selected) representation of its data;
4. ANANSI is connected to at least one non-CLARIAH LOD-enabled data repository and serves a (selected) representation of its data;
5. ANANSI is accessible for machines through a web API;
6. ANANSI is accessible for human users through a front-end web environment;

The LOD-enabled data repositories in WP3, 4 and 5 need to be available. This is an external dependency. Meertens Institute is responsible for the LOD enabled repository in WP3, IISH is responsible for the LOD enabled repository in WP4 and NISV is responsible for the LOD enabled repository in WP5.

Secondly, ANANSI will be considered successful when it meets a set of metrics defined by the development team early in the project. These metrics may include:

1. a number of humanities research questions that can be answered X% faster;
2. a number of external humanities software systems/users that regularly connect to ANANSI to perform operations;
3. a number of daily operations;

Thirdly, ANANSI will be considered successful when it is accessible through a number of external software systems frequently used in (digital) humanities research. The selection of these software systems will be the responsibility of the development team after due inventorization of user wishes, requirements and demands. E.g.:

1. research platforms like Gephi, Paladio, GATE, NodeGoat etc.
2. programming libraries and packages for Python, R, Java etc.

Deployment

Preliminary

At this early stage only a preliminary description of the expected deployment can be given. ANANSI will likely be an implementation of JSON-LD powered by either a native triple store or a more generic graph database solution. The data will be made accessible through a web API that is likely REST-based. A query endpoint will be part of the solution. This is unlikely to be an open SPARQL-endpoint because of expected issues of scale. The querying API is more likely to be based on industrial indexing services like SOLR or ElasticSearch. The back-end infrastructure will be developed in JAVA.

Software Release Environment

The team develops ANANSI in short iterative sprints aimed at delivery of working software. These sprints are presided over by the product owner. Responsibility for the practical deployment and the selection of software solutions rests with the team.

The front-end environment will be developed in JavaScript/REACT and run natively in a web browser. It will probably implement existing Huygens ING libraries for faceted search and editing forms.

The software release environment will be hosted by Huygens ING and aim for continuous integration based on a Maven, Jenkins toolchain with unit, integration and acceptance tests.